基于 VAE 对软测量中缺失数据的填补方法

李应聪

(北京化工大学信息科学与技术学院)

摘要： 软测量在数据进行采集到应用的途中经常会出现数据缺失的情况，这将会大大降低建立的模型的精度。本文提出了基于变分自编码器(Variational Autoencoder)与 GRU 神经网络建立的填补模型，并通过实际工业流程验证了填补后数据的准确性。最后实验表明 VAE 填补模型在缺失率为 10%和 30%情况下的 RMSE 和 MAE 分别为 3.316%，4.262%和 2.386%，2.964%，相较于其他填补算法 PCA 和 SVD 均有更显著的效果，验证了该模型的可行性。

关键字： 缺失数据填补　GRU　VAE　软测量

Filling Missing Data in Soft Sensing based on VAE

Yingcong Li

(School of Information Science and Technology, Beijing University of Chemical Technology)

Abstract: In the realm of soft sensing, missing data frequently occurs during the journey from data collection to application, significantly diminishing model accuracy. This paper introduces a filling model based on the Variational Autoencoder (VAE) and GRU neural network. Validation through industrial processes confirms the accuracy of the imputed data. Experimental results demonstrate that the VAE imputation model yields an RMSE and MAE of 3.396% and 2.458% for missing rates of 10%, and 3.549% and 3.078% for missing rates of 30%, respectively. Compared to alternative imputation algorithms like PCA and SVD, the VAE model exhibits significantly enhanced performance, affirming the feasibility of this model.

Keywords: Missing data filling　GRU　VAE　Soft sensor

引言

在如今的工业过程中需要实现对工业过程的实时监控，质量控制，异常检测，生产优化等目的，都是通过对工业过程在的数据进行软测量建模[1]和分析实现，而这一过程就是软测量[2][3]。然而数据缺失在软测量中是很常见的问题，在实际应用中常常会由于各种原因，如传感器故障，数据采集错误等，过程数据就可能会出现缺失[4][5]，而这些数据的缺失就会严重影响软测量的建模和预测精度，进而影响到工业生产质量，甚至影响到生产的安全。因此软测量研究和应用中，填补缺失数据，提高模型的鲁棒性和预测能力至关重要。

目前数据填补方法分为三种，第一种是将缺失的数据直接删除，但是这样的操作无疑会造成数据中重要信息的缺失，在实际应用中基本上不会采用。第二种是基于数学统计的方法[5]，早在 20 世纪 50 年代，缺失数据的填补就在统计学领域中被提出，当时的人们主要采用简单的插值法，例如均值法，中位数法，最近邻法等，这些方法虽然简单，但是忽略了数据间的相关性以及时序特征，因此填补的准确度并不高。第三种是基于机器学习的方法[7][8]，例如 K 近邻[9]，最大似然估计，矩阵分解理论[10]等，这些方法能够进一步的学习到数据间的相关性，并且模型构建简单，计算参数也较少。但是在工业生产过程中，生产机理复杂，并且数据往往会受到多种因素的影响，而数据的分布不尽相同，数据之间有强相关性，并且在时间维度上也有强烈的上下文关系，这些传统的机器学习方法假设的分布与实际分布不一定符合，并且这些模型往往忽视了数据的时序特征，所以以上的方法均不能很好处理生产过程中数据填补的问题。

近年来，深度学习方法在缺失数据填补邻域广泛应用。变分自编码器（VAE）是一种基于深度学习的生成模型，是在2013年由 Kingma 和 Welling 提出的[11]，该模型由于其能够学习数据的潜在特征表示，生成符合真实样本分布规律的数据，目前广泛应用了各种数据样本生成领域。文献[12]研究了 VAE 在处理异构数据的应用，在 VAE 的基础上，通过对连续数据和离散数据独立设计神经网络，提出了一种可以处理异构数据的模型。这些模型结构简单，模型训练效率高，说明了 VAE 模型能够很好的学习到数据间的特征。文献[13]将 VAE 用于人脸修复生成，采用变分自编码器（VAE）对输入的基准人像进行低程度的人脸重构，然后将真实的基准人像和低清晰度重构人像通过隐空间转换进行域对齐，在该隐空间中学习人像修复，在保留人像的基本构图后重新得到一张全新的高清细节人像，验证了 VAE 模型生成数据的准确性。GRU（Gated Recurrent Unit）是一种循环神经网络模型，于2014年由 Cho 等人提出，该网络能够很好的学习到时间序列的时序特征，并对数据进行动态预测，但是对于学习数据间的潜在特征能力还有待提高。

综上所述，现有模型不能满足工业生产中缺失数据填补的精度需求。所以本文利用 VAE 提取单一序列的自相关性信息和多维数据的互相关信息获取不同在线检测状态量数据的相关性，再结合 GRU 神经网络学习数据分布，挖掘时序特征，提出 VAE-GRU 缺失数据填补模型，并通过实际脱丁烷塔工业过程在线检测数据对模型的理论有效性和工程适用性进行验证。

1　模型研究

1.1　VAE 模型

变分自编码器拥有与自编码器结构相似的编码器和解码器以外，其中编码器的作用是学习输入数据的潜在特征的空间表示，而解码器的作用是从获得到的降维特征中重构出原始的输入数据，还额外增加了隐含变量，可以看作是贝叶斯网络和神经网络的混合体，它的实质就是将提取到的每个潜在特征表示为概率分布，当从潜在状态解码时将从每一个分布中进行随机抽样，生成一个特征向量作为解码器的输入，这就是变分自编码器模型最精妙的地方，通过将特征表示为概率分布而不是像自编码器那样简单的特定编码表示，使得变分自编码器有了强大的生成能力，甚至能够生成原本的样本中原来没有的新样本。在变分自编码器模型中编码器实质就是一个变分推断网络，它的功能就是用来得到输入样本的均值和方差，而变分自编码器模型的解码器则可以看为是一个将隐含变量映射到原始变量的生成网络。它的具体结构如图 1 所示：

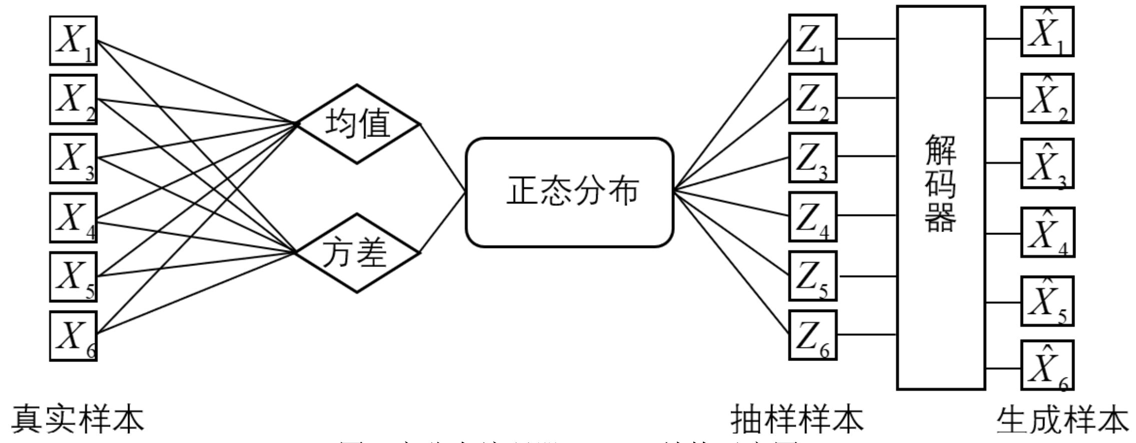


图 1 变分自编码器（VAE）结构示意图

通过图 1 可以清晰的看到变分自编码器共分为 3 个部分：

（1） 编码层：真实样本通过编码层以后推断出了潜在分布的方差和均值

（2） 隐含变量：也就是图 1 中通过均值和方差得出的属于真实样本的专属正态分布

（3） 解码层：对隐变量采样以后得到特征向量，解码层通过该特征向量来生成新的样本

该模型能够很好的学习到数据的潜在表示和概率分布，从而能够更好地理解数据的结构和特征，这使得该模型填补的数据能更加符合实际的情况，这是选择该模型作为填补模型的最大原因，因为实际的工业过程数据必须是符合实际情况，这样才能保证模型预测的准确性。其次该模型还可以处理非线性数据，进行非线性的建模，有着处理复杂数据的能力，这使得该模型更加适用于工业中。但是该模型也存在着不足的地方，就是该模型在学习数据的潜在表示和生成新的数据时，通常假设数据是独立同分布的，并且不考虑时间序列数据的动态性，这就意味着传统的 VAE 模型不能考虑到工业过程中的动态性和序列性。这个不足将导致该模型的建模预测效果不达标，所以针对这个问题本文给出了相应的解决方案即对该模型进行改进。将 VAE 模型和递归神经网络（RNN）结合起来，以考虑时间序列数据的动态性和变化规律，本文选择的 RNN 网络是 GRU 神经网络。经过改进得到的 VAE-GRU 模型可以在生成过程中考虑过去的状态和当前的观测，从而实现更加准确的预测和生成。

1.2 GRU 网络

GRU 神经网络的节点结构相比于 RNN 来说复杂了很多，除了状态信息和当前时刻输入之外多出了很多环节，原因在于 GRU 网络引入了门控机制，所谓门控机制实质就是控制信息流动和遗忘的机制。GRU 网络引入了两个门，分别的重置门和更新门，如图 2 方框 1 和方框 2 部分所示，分别对两个门的作用以及原理进行介绍：

（1） 重置门：重置总的来说就是控制历史信息的遗忘，用于更新当前时刻的隐藏状态。也就是控制前一个状态的隐藏信息h_{t-1}哪些能传进候选状态，通过 sigmoid 激活函数来控制当前时刻输入和上一时刻隐藏状态的重要程度，通过点乘运算将当前时刻输入和上一时刻隐藏状态的输出相乘，得到一个加权后的输入值，用于更新当前时刻的隐藏状态。

（2） 更新门：更新总的来说就是控制当前信息的选择性传递，也就是控制前一状态的信息h_{t-1}有多少信息能够保留到新状态h_t中。通过 sigmoid 激活函数来控制当前时刻隐藏状态的更新程度，通过点乘运算将当前时刻输入和上一时刻隐藏状态的输出相乘，并加上一项（1-更新门控单元的输出）与当前时刻隐藏状态的点乘运算，得到一个加权后的输出值，作为当前时刻的隐藏状态。

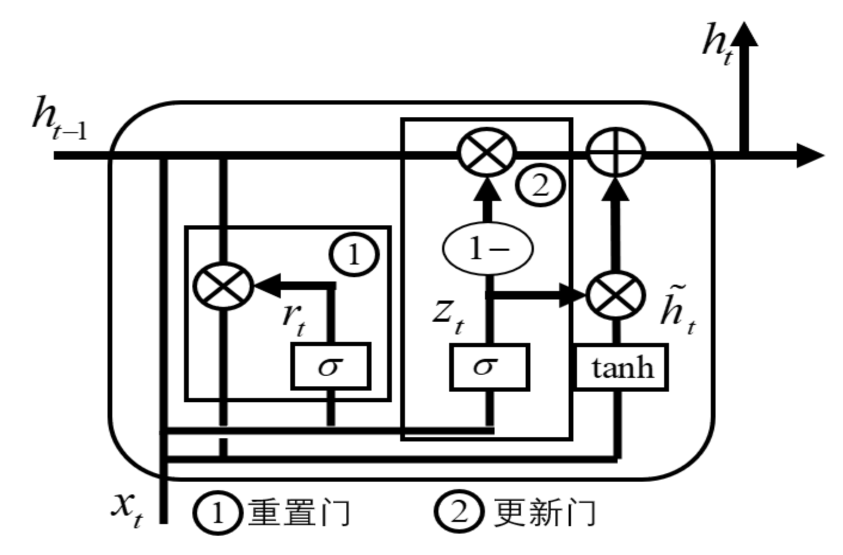


图 2 GRU 单元节点结构示意图

实际操作如公式 1-4 所示：

$$z_t = sigmoid(W_z[h_{t-1}, x_t]) \qquad (1)$$

$$r_t = sigmoid(W_r[h_{t-1}, x_t]) \qquad (2)$$

$$\tilde{h}_t = tanh(W[r_t \odot h_{t-1}, x_t]) \qquad (3)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \qquad (4)$$

经过上面控制门操作之后，也就是进行了更新记忆操作，能够调整前一状态的信息对当前时刻的状态的影响，从而有选择性地保留重要信息，抑制不重要信息，这种选择性的保留和抑制便能够有效地缓解梯度消失和梯度爆炸的问题

2　VAE-GRU 模型构建

传统的 VAE 模型不能考虑到工业过程中的动态性和序列性，所以为了解决这个问题提出了一种将 GRU 神经网络与传统 VAE 模型相结合的解决方案。因此将 GRU 网络加入到传统的 VAE 模型中，能够得到新的 VAE-GRU 模型，该模型能解决工业过程中动态性的问题。之所以 GRU 能和 VAE 模型结合在一起是因为 GRU 网络的实质其实就是一层特殊的神经网络，而传统 VAE 模型的编码器和解码器是由线性神经网络构成的全连接层，所以只需要将 GRU 神经网络分别加入到 VAE 模型的编码器和解码器当中就可以实现两者的结合。该模型的结构如图 3 所示：

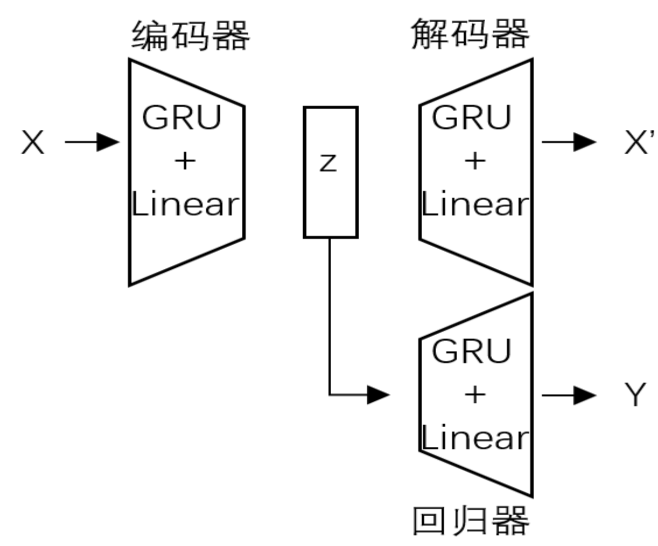


图 3 VAE-GRU 结构示意图

如图 3 所示，只需要传统 VAE 模型中的编码器和解码器中加入 GRU 神经网络层即可完成两者的结合，GRU 网络能帮助 VAE 更好的提取输入的特征，并使得该模型能够关注到数据序列中的前后依赖关系即工业流程的动态性，其次在 VAE 模型中加入回归器模块可以实现数据建模并输出对数据序列的预测值。

3　实验设置与结果分析

3.1　数据说明

为了验证该模型的实用价值和有效性，使用到了脱丁烷系统过程建模中，用于预测脱丁烷塔底部丁烷含量。脱丁烷塔是炼油厂中重要的分离设备之一，用于将石油中的丁烷和丁烯分离出来，在脱丁烷塔的生产过程中，各种因素会影响其操作效率和产品质量，如温度，压力，流量等参数变化。图 4 具体的描述了这个过程，塔底丁烷含量作为衡量脱丁烷效果的关键过程变量需要被重点关注。传统的丁烷量观测采用物理模型和经验模型，需要大量的人工干预和参数调整，且在处理非线性，多变量，高维数据时存在一定的局限性，并且具有较大的滞后性。针对这些问题一般采用软测量技术，可以更好的帮助优化工艺参数，提高产品质量和能源利用率，实现更加高效，更精准，更自动化的生产控制。

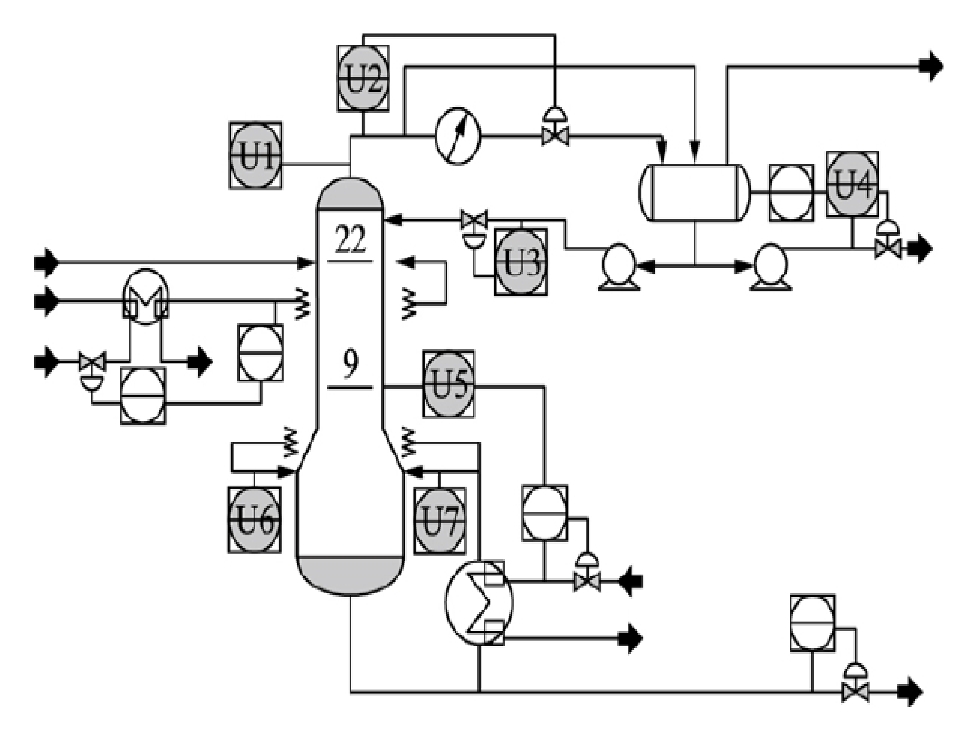


图 4 脱丁烷塔过程流程图

经过一定的相关性分析之后，为了更加精确控制脱丁烷塔工作效率，选取出了 7 个与塔底丁烷含量相关性及强并且可以即时观测的过程变量，用于建立软测量模型。表 1 给出了这些变量的具体描述以及具体数据，主要包括温度，压力，流量等。这些数据一共为 2394 组，每组七个变量，在进行实验室按照实验设定将数据集分为训练集和测试集，然后再进行实验。

表 1 脱丁烷塔过程变量描述表

序号	过程变量
1	炉顶温度(℃)
2	炉顶压力(kPa)
3	回流流量(m³.s-1)
4	流向下一个过程的流量(m³.s-1)
5	第 6 隔板温度(℃)
6	炉底温度 A(℃)
7	炉底温度 B(℃)
8	底部丁烷含量(%)

3.2 实验设置与步骤

3.2.1 实验设置

实验采取对比试验的方式,采用多种不同的填补方式对不同缺失率的不完整数据进行填充,然后分别将每个算法填充好的数据放到训练好的建模预测模型中

进行预测，最后输出各个算法的均方根误差进行评价，并给处可视化追踪图和误差盒图。实验中涉及的算法包括传统 VAE 模型填补，VAE-GRU 模型填补，PCA 填补 3 种方法。其中采用的评价指标是均方根误差（RMSE），它的实质就是用来测量数据之间差异的量度，在这里是用来计算实验中预测值与真实值之间的误差均方差误差的计算公式 5 所示：

$$R_{rmse} = \sqrt{\frac{1}{N} \times \sum_{i}^{N} (X_i - \hat{X}_i)^2} \tag{5}$$

上式中N表示缺失数据的个数，X_i和$\hat{X}_i$分别表示真实数据和数据预测值。

3.2.2 实验步骤

首先搭建完整 VAE-GRU 模型，该模型主要的功能是实现对完整数据集进行回归建模，然后输出数据的预测值。使用不含缺失值的数据对模型进行训练，由回归器建模产生预测值，利用生成的预测值和训练值的均方误差以及 VAE 生成器的重构损失反向传播，优化模型建模预测的真实性，实现对完整数据的建模预测。模型的搭建描述如下：

（1） 设定模型相关参数，包括神经网络的层数，神经元个数等相关参数，并对全体样本进行处理时序化处理，样本划分，归一化处理

（2） 使用训练数据集对模型进行训练，用梯度下降法和反向传播法来实现整个网络的训练学习，并计算出均方根误差和 KL 散度

（3） 将测试数据集放入训练好的模型当中，进行建模预测，并给出最后的预测误差和可视化追踪图

（4） 将该模型保存用于经过填补后的完整数据的建模预测

然后搭建 VAE 填补模型，VAE-GRU 填补模型，PCA 填补模型，分别将各个模型用于含缺失值数据的填补，然后将填补好的数据用于回归预测，观察各组填补数据得到预测值与真实值的差异，最后分析各个填补模型。

3.2.3 模型结构

结构设计方面，将 GRU 网络加入到 VAE 的编码器和解码器结构中，其中编码器输出，解码器输出和回归层输出均采用 LeakyRelu 激活函数。其它网络充分考虑模型计算量设置如表 2 所示：

表 2 VAE-GRU 模型参数设置

所属模型结构		节点数
编码层	GRU	12
编码层	线性层	20
编码层	线性层	20
解码层	GRU	6
解码层	线性层	16
解码层	线性层	12
回归层	GRU	6
回归层	线形层	12

3.2.4 模型对比分析

三个模型分别对数据集缺失率为 10%和 30%情况下进行填补，将填补后的数据回归预测得到实验结果如图 5-图 10 所示：

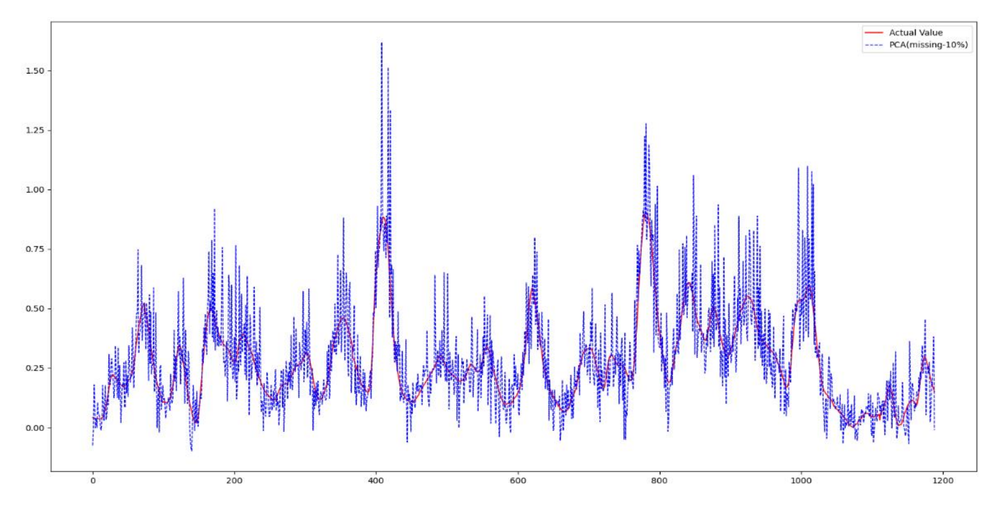


图 5 PCA 算法填补建模结果（缺失率 10%）

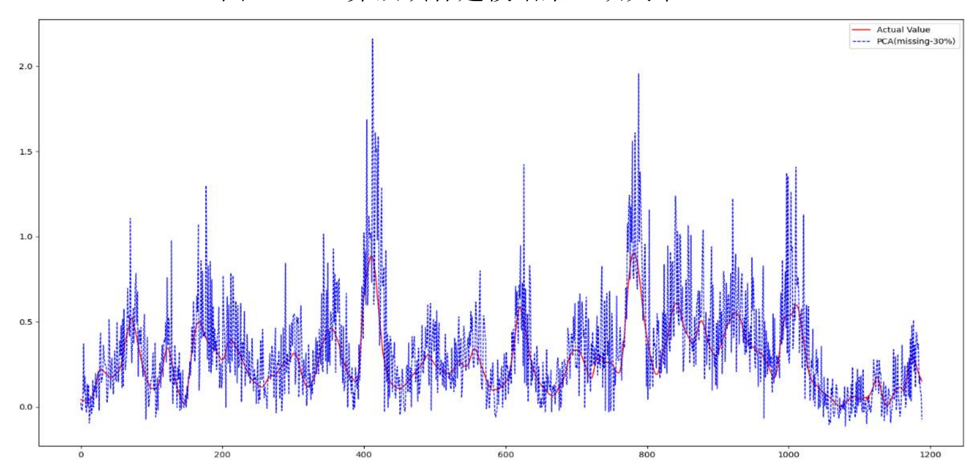


图 6 PCA 算法填补建模结果（缺失率 30%）

从结果看出不同缺失率对 PCA 算法的填补有较大影响，并且整体看 PCA 的填补精度很不理想，这是因为 PCA 算法在进行填补时是假设数据应该服从线性分布，但是在实际工业流程中，数据往往是呈非线性和非正态分布，这会导致处理数据时失去一些信息。其次 PCA 是一种无记忆的算法，也就是它将每个时间步的样本视为独立的数据点，并没有考虑到时间序列数据的先后顺序和时间关系。

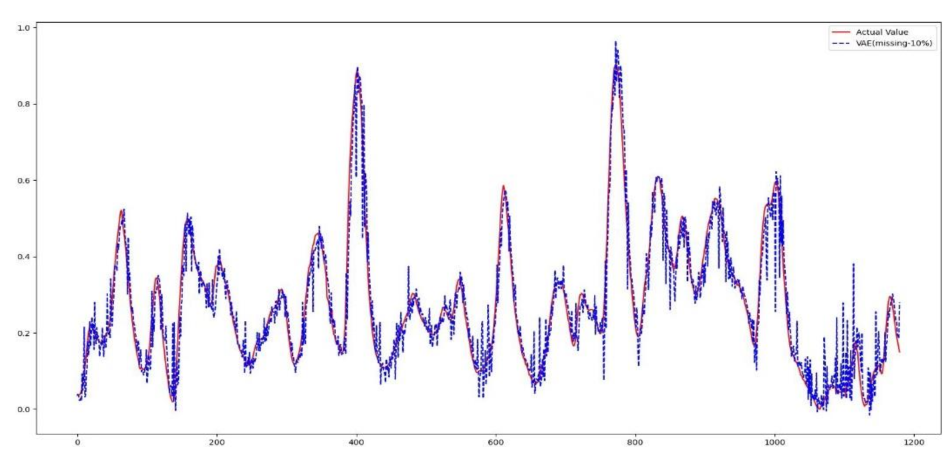


图 7 传统 VAE 模型建模预测结果（缺失率 10%）

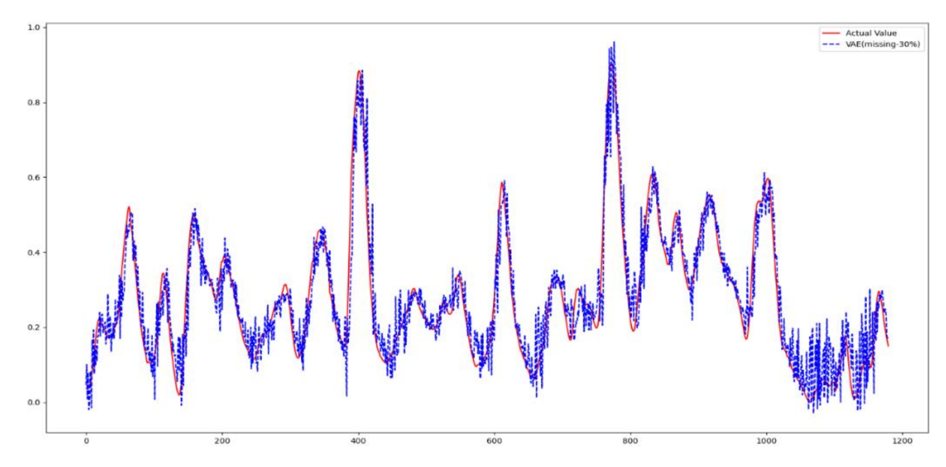


图 8 传统 VAE 模型建模预测结果（缺失率 30%）

从上述结果可以直观地看到传统 VAE 模型对时间序列数据的填补效果并不好，不管是 10%缺失率还是 30%缺失率误差都在 5%以上，可以说建模效果不理想。之所以出现这样的结果是因为传统 VAE 模型在填补数据时不能很好的考虑到工业过程数据的动态性，所以预测的结果跟实际工业流程的结果会有较大偏差，并且当缺失率增大时对模型的准确度影响很明显，从可视化追踪图中可以很清晰的看出来，缺失率在 30%的时候误差明显增大了不少

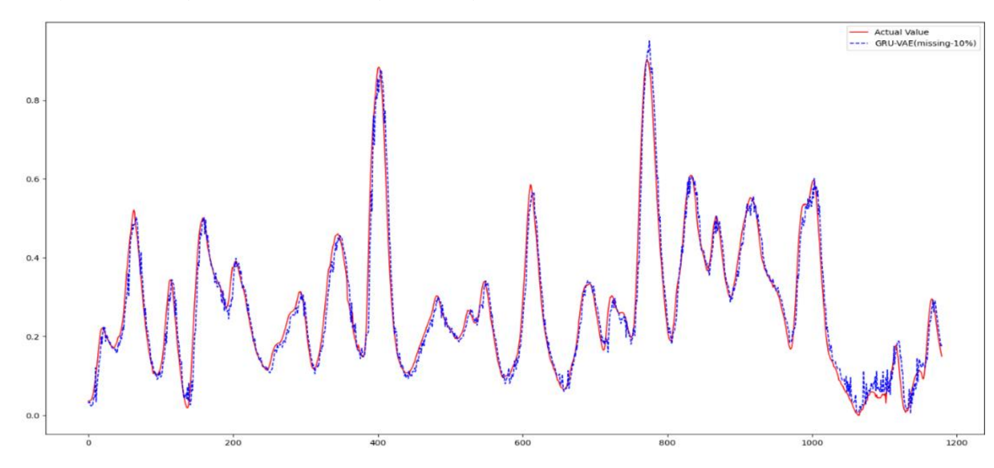


图 9 VAE-GRU 模型建模预测结果（缺失率 10%）

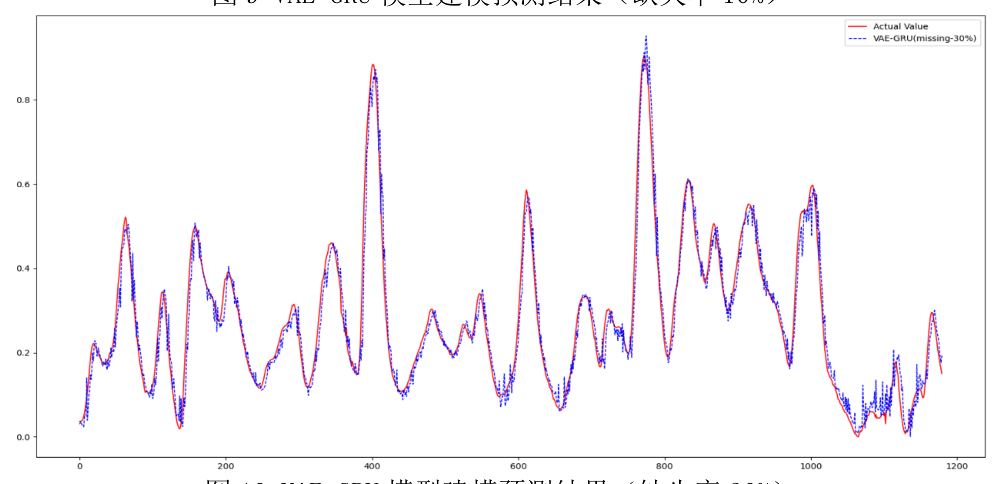


图 10 VAE-GRU 模型建模预测结果（缺失率 30%）

从上述结果看出，经过 GRU 网络改进的 VAE-GRU 模型在实验中的表现有

显著提高。首先是准确度方面，误差有了明显降低，已经降到了 5%以内，这已经能够很好的说明该模型的有效性和该改进的可行性。其次模型的稳定性方面也有了很好的提高，传统的 VAE 模型在缺失率不同的情况下，模型准确度会有较大的影响，但是改进后的 VAE-GRU 模型在不同缺失率下准确度非常稳定。所以经过以上两个模型的结果对比以后能很好的证明该模型对传统 VAE 模型确实能有很大的提升。

表 3 不同模型预测对比效果

模型	RMSE(10%)	RMSE(30%)	MAE(10%)	MAE(30%)
传统 VAE	4.020%	5.904%	2.884%	4.341%
PCA	12.724%	21.468%	8.713%	15.121%
VAE-GRU	3.316%	4.262%	2.386%	2.964%

很明显可以看出不管在多少数据缺失率下，三个模型中 VAE-GRU 模型的预测误差都是最小，并且相比较于传统 VAE 模型在准确度上也有了不小的提升。这充分验证了将 GRU 网络和 VAE 相结合的理论可行性，说明 GRU 网络能够 VAE 模型带来更好的收益，也说明了 VAE-GRU 模型有更好的模型拟合能力，更强的预测能力。

4 结论

针对数据缺失这种情况，提出了一种基于变分自编码器（Variational Auto encoder，VAE）的填补模型，该模型能够很好的将数据进行降维处理并获取数据中的隐藏特征，再根据得到的数据特征给出缺失数据的预测值，进而能使用预测值将缺失的数据填补上。但是又考虑到 VAE 模型不能考虑到工业过程中的动态性和序列性，本文提出了将递归神经网络与传统 VAE 模型相结合的解决方案。递归神经网络选择到了 GRU 神经网络，基于该神经网络的良好性能与传统 VAE 模型相结合，构成的 VAE-GRU 模型不仅能够增强对数据特征的提取能力，还能考虑到数据序列的动态性和变化规律，并且该模型拥有很强的建模能力，VAE 部分负责潜在特征的提取，GRU 部分则负责建模时序数据的动态演变。

基于以上的理论，建立了 VAE-GRU 模型，并应用于脱丁烷塔中关键过程量塔底丁烷量的预测，同时与目前常用的缺失数据填补算法以及传统 VAE 模型进行了对比试验，在实验中本文建立的模型取得了最好的结果，验证了模型的有效性和实用性。在 10%数据缺失情况下，VAE-GRU 模型的输出结果均方根误差为 3.316%，平均误差为 2.386%，在 30%缺失率情况下，输出结果均方根误差为 4.262%，平均误差为 2.964%，相比于其它四种填补算法的表现来看还是很不错的。

当然，本文建立的模型还是存在一些问题和不足，例如该模型不能自动适应多种维度的数据，每次使用不同的数据都需要先对模型进行训练，并且需要对数据进行不同的预处理。另外，本文所建立的模型在缺失率较高的情况下准确度暂时还不是非常高，仍存在一定误差，暂时还没有找到较好的方式来降低这个误差，这些问题都有待于今后进一步探索和研究。

参考文献：

[1] 徐宗煌.基于多元线性回归分析的汽油辛烷值损失预测建模[J].宁夏大学学报(自然科学版),2022,43(01):22-29.

[2] 胡长松,贾延刚,夏伯楷.软测量技术现状与发展[J].石油仪器,2003,17(3):4-6.

[3] 李修亮. 软测量建模方法研究与应用[D]. 浙江大学, 2009.

[4] ZHANG N H.Methodolgical progress note：handling miss-ing data in clinical research[J].Journal of Hospital Medi-cine，2020，15（4）：237-239.

[5] GOMILA R，CLARK C S.Missing data in experiments：challenges and solutions[J].Psychological Methods，2020.

[6] KAISER J. Dealing with missing values in data[J]. Journal of Systems Integration, 2014, 5(1): 42-51.

[7] MIRANDA V, KRSTULOVIC J, KEKO H, et al. Reconstructing missing data in state estimation with autoencoders[J]. IEEE Transactions on Power Systems, 2012, 27(2): 604-611.

[8] MAZUMDER R, HASTIE T, TIBSHIRANI R. Spectral regularization algorithms for learning large incomplete matrices[J]. The Journal of Machine Learning Research, 2010, 11: 2287-2322.

[9] ZHANG S C. Nearest neighbor selection for iteratively kNN imputation[J]. Journal of Systems and Software, 2012, 85(11): 2541-2552.

[10] HASTIE T, MAZUMDER R, LEE J D, et al. Matrix completion and low-rank SVD via fast alternating least squares[J]. The Journal of Machine Learning Research, 2015, 16(1): 3367-3402.

[11] Kiingma D P,Welling M. Auto-encoding variational bayes. arXiv preprint arXiv: 1312.6114, 2013.

[12] Alfredo Nazábal, Pablo M. Olmos, Zoubin Ghahramani, Isabel Valera,Handling i-ncomplete heterogeneous data using VAEs,Pattern Recognition,Volume 107,2020,107501,ISSN 0031-3203,

[13] 丁一鸣,黄晨,孔聪,孔祥懿,庞毅林.一种基于VAE的人脸修复式生成方法[J].网络安全技术与应用,2022(05):47-49.